A VECTOR STATISTICAL PIECEWISE POLYNOMIAL APPROXIMATION ALGORITHM FOR ENVIRONMENT COMPENSATION IN TELEPHONE LVCSR

Zhaobing Han, Shuwu Zhang, Huayun Zhang, Bo Xu

National Lab of Pattern Recognition, Institute of Automation
Chinese Academy of Sciences, Beijing, China, 100080
{zbhan, swzhang, hyzhang, xubo }@hitic.ia.ac.cn

ABSTRACT

In the paper, a Vector statistical Piecewise Polynomial (VPP) approximation algorithm is proposed for environment compensation that speech signals are degraded by both additive and convolutive noises. By investigating the model of the telephone environment, we address a piecewise polynomial, namely two linear polynomials and a quadratic polynomial, to approximate the environment function precisely. The VPP is applied either to the stationary noise, or to the non-stationary noise. In the first case, the batch EM is used in log-spectral domain; in the second case the recursive EM with iterative stochastic approximation is developed in cepstral domain. Both approaches are based on the minimum mean squared error (MMSE) sense. Experimental results are presented on the application of this approach in improving the performance of Mandarin large vocabulary continuous speech recognition (LVCSR) due to the background noises and different transmission channels (such as fixed telephone line and GSM). The method can reduce the average character error rate (CER) by about 18%.

1. INTRODUCTION

Basically, the distortion sources in telephone network fall into two categories: (1) noise contamination including background noise and electrical noise, and (2) channel effect caused by transmission line and telephone handset. Due to these distortion sources, the speech recognition performance will be seriously damaged [1].

In the recent years, many methods have been proposed for compensating the noisy effect. Accordingly, the Codeword Dependent Cepstral Normalization method (CDCN) [2], the Multivariate Gaussian Based Cepstral Normalization (RATZ) and a Vector Taylor Series (VTS) [3] were presented for reducing the variability of additive noise and channel effect. These approaches use the MMSE estimator to model the noisy speech as the clean signal plus a correction vector. In CDCN, the correction vector is a weighted sum of codeword-dependent correction. However, the noise only depends on the first codebook and the correction vector is constant for all vectors in a codebook. Although the correction of RATZ is a weighted sum of multivariate Gaussian mixture and it doesn't have the above two disadvantages, the correction doesn't represent the inner structure of the noise environment. VTS uses an environment function to describe the correction vector in detail and analytically approximates the function by the truncated Taylor series expansion around the mean of the clean speech. But, it is known that Taylor series is not a precise approximation if the distribution of the variable isn't around the point of Taylor's expansion.

In this paper, we propose a vector piecewise polynomial method, in which we approximate the environment function piecewisely (two linear polynomials and a quadratic polynomial). By taking account of the property of the environment function, the approximation by using VPP method is very close to the actual function. Furthermore, the approach doesn't so strongly rely on the expansion point as VTS. In addition, an HMM with Gaussian outputs is exploited to classify the acoustic space of clean speech.

The paper is organized as follows: First the property of the environment function is analyzed. We then explain how to use the piecewise polynomial to approximate it. After that, the environment estimation framework of the stationary and non-stationary noise using VPP is introduced. And finally, the experimental results are given and some experimental conclusions are drawn.

2. VPP APPROXIMATION ALGORITHM

2.1. A Model Of The Environment

As we all know, a commonly accepted model in the log-spectral domain is:

$$y = x + h + \log(1 + e^{n-x-h}) \qquad (1)$$

Or in more general terms:

$$y = x + f(x, n, h)$$

Where **h** is an unknown parameter that represents the effect of linear filter and **n** represents the effect of background noise. f(x,n,h) denotes the environment function.

0-7803-7663-3/03/$17.00 ©2003 IEEE

To simplify the notation, we define the vector function g(v) as

$$g(v)=\log(1+\exp(v)) \qquad (2)$$

g(v) is a nonlinear function. VTS uses the Taylor series expansion around the operating points (μ_x , n_0 and h_0) to approximate the g(v). The accuracy of this approximation strongly depends on the initial values or the operating points.

2.2. The Property g(v)

From Fig.1, g(v) is monotonically increasing, with asymptotes at g(v)=0 for $v \to -\infty$ and g(v)=v for $v \to +\infty$. In the ranges $(-\infty$, a) and (b,$+\infty$) the plot of g(v) is very close to a line; in the range (a , b) , the plot is a curve. In our experiment it proves that a quadratic polynomial is enough to approximate g(v) in the range (a, b), and doesn't need a cubic or higher order polynomial. Therefore, we adopt a quadratic polynomial to approximate the curve.

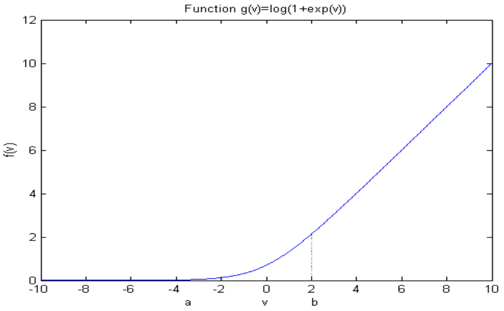

Fig.1 the function g(v)=log(1+exp(v))

2.3. The Piecewise Polynomial Approximation

In our approach, linear approximation in the range of $(-\infty$,a) and (b, $+\infty$) is obtained by minimizing the mean squared error. The range of (a, b) is replaced by a quadratic Chebyshev Polynomial.

The function prototype of the linear function is:

$$g(v) = Av + B$$

And the quadratic function is:

$$g(v) = Av^2 + Bv + C$$

Where A, B and C are constant and determined beforehand.

In previous work [4], it was shown that it is reasonable to adopt the Gaussian assumption for clean and noisy speech. This assumption implies a linear transformation between the log spectra of clean and noisy speech for each Gaussian density component of the PDF of clean speech. Therefore, to estimate the noise and channel parameters, we approximate the environment function (Eq.(1)) by the statistical linear transformation:

$$\tilde{y} = (1-A_k)x + (1-A_k)h + A_k n + B_k \qquad (3)$$

So the estimate of mean and variance of the Kth Gaussian of the noisy speech can be gotten as functions of A_k , B_k :

$$\mu_{y,k} = (1-A_k)(\mu_{x,k} + h) + A_k \mu_n + B_k \qquad (4)$$

$$\Sigma^2_{y,k} = (1-A_k)^2 \Sigma^2_{x,k} + A^2_k \Sigma^2_n \qquad (5)$$

Since the estimation of $\mu_{y,k}$ and $\Sigma^2_{y,k}$ can be obtained from the VPP algorithm. From (4) and (5), we can compute A_k and B_k . Therefore, μ_n , $\sum_n$ and **h** will be derived from EM algorithm. After that, we use the MMSE criterion to calculate the clean speech given the observed noisy speech.

Compared with VTS, VPP makes use of the HMM parameters and doesn't strictly depend on the initial values. The Vector Polynomial approximationS (VPS) algorithm (B.Raj, 1996) is analogous to VPP. But VPS uses the triangular function to approximate the second derivative of environment function. Although it may be very exact to approximate the derivative, its approximation of the function g(v) which is derived from integrating the derivative twice may be not precise. Moreover, it adds the computational complexity.

3. THE ENVIRONMENT ESTIMATION FRAMEWORK

In this section, we first apply the VPP algorithm to estimate the stationary noise and telephone channel with the batch EM. Then the based VPP estimation of the non-stationary noise and channel using the recursive EM is proposed. The formulas of both estimation methods are derived, respectively.

3.1. The Estimation of Stationary Noise and Channel

In this case, the optimal parameter is $\lambda = \{\mu_n, \Sigma_n, \text{h}\}$.

The auxiliary function, Q, is defined by

$$Q(\lambda,\overline{\lambda}) = E[\log p(X,N,K \mid \overline{\lambda} = \{\mu_n,\Sigma_n,h\}) \mid y, \lambda = \{\mu_n,\Sigma_n,h\}]$$

Where **N**={**n**$_1$,**n**$_2$,….,**n**$_T$} denotes the noise vector sequence which is statistically independent of the clean feature vector sequence X in log-spectral domain, and K={k$_1$,k$_2$,….,k$_T$} is a hidden sequence of mixture components. By following the way similar to that used in [5], we can obtain the formulas of the estimation of the environment parameters:

$$\widehat{\mu}_n = \frac{\sum_{t=1}^{T}\sum_{k=1}^{K} P(k \mid y_t,\lambda)\mu_n(y_t,k,\lambda)}{\sum_{t=1}^{T}\sum_{k=1}^{K} P(k \mid y_t,\lambda)}$$

$$\widehat{\Sigma}^2_n = \frac{\sum_{t=1}^{T}\sum_{k=1}^{K} P(k \mid y_t,\lambda)[\Sigma_n(y_t,k,\lambda) + \mu_n(y_t,k,\lambda)\mu_n^T(y_t,k,\lambda)]}{\sum_{t=1}^{T}\sum_{k=1}^{K} P(k \mid y_t,\lambda)} - \hat{\mu}_n\hat{\mu}_n^T$$

$$\hat{h} = \sum_{t=1}^{T}\sum_{k=1}^{K} P(k \mid y_t, \lambda)((I-A)^T)^{-1} \Sigma_{x,k}^{-1}(y_t - (I-A)\mu_{x,k} - An - B)$$

Where

$$\mu_n(y_t, k, \lambda) = \tilde{\Sigma}_n(\Sigma_n + \tilde{\Sigma}_n)^{-1}\mu_n + \Sigma_n(\Sigma_n + \tilde{\Sigma}_n)^{-1}\overline{\mu}_n$$

$$\Sigma_n(y_t, k, \lambda) = [\Sigma_n^{-1} + \tilde{\Sigma}_n^{-1}]^{-1}$$

Where

$$\overline{\mu}_n = (A^T)^{-1}(y_t - (I-A)\mu_x - (I-A)h - B]$$

$$\tilde{\Sigma}_n = (A^T)^{-1}(I-A)^T \Sigma_{x,k}(I-A)(A^T)^{-1}$$

3.2. The Estimation Of Non-stationary Noise and Channel

Now, the compensation is applied in cepstral domain. So the environment model is derived from (1):

$$y = x + h + C\log(I + e^{C^T(n-x-h)}) \qquad (6)$$

Where C is the discrete cosine transform matrix.

Recursive noise parameter estimation is a solution to the recursive EM optimization problem [6] [7] [8].

$$n_{t+1} = \arg\max_n Q_{t+1}(n).$$

The objective function $Q_{t+1}(n)$ above is the conditional expectation.

$$Q_{t+1}(n) = E[\ln p(y_1^{t+1}, K_1^{t+1} \mid n) |, y_1^{t+1}, n_1^t]$$

Where K_1^{t+1} = k_1 , k_2 ,…., k_{t+1} is the sequence of (hidden) mixture components in the clean speech model up to time t+1.By following the way similar to what is used in [8], we have:

$$Q_{t+1}(n) = \varepsilon \cdot Q_t(n_t) - R_{t+1}(n_{t+1}) \qquad (7)$$

Where

$$R_{t+1} = \sum_{k=1}^{K} \gamma_{t+1}(k)[y_{t+1} - \mu_k^y(n_{t+1})]^T(\Sigma_k^y)^{-1}[y_{t+1} - \mu_k^y(n_{t+1})]$$

The forgetting factor ε is based on a tradeoff between the strength of noise tracking ability and the reliability of noise estimate. The occupancy probability $\gamma_\tau(k)$ is computed using Bayes rule in the E-step.

We can prove that $Q_{t+1}(n_{t+1})$ in recursion Eq.7 is maximized via the following recursive form of noise parameter updating:

$$n_{t+1} = n_t + D_{t+1}^{-1} s_{t+1} \qquad (8)$$

Where

$$\mathrm{s}_{t+1} = \frac{\partial \mathrm{R}_{t+1}}{\partial \mathrm{n}}\Big|_{n=n_t} = \sum_{\mathrm{k}=1}^{\mathrm{K}} \gamma_{t+1} A^T(\Sigma_k^y)^{-1}(y_{t+1} - \mu_k^y(n_{t+1}))$$

$$D_{t+1} = \frac{\partial^2 Q_{t+1}}{\partial^2 n}\Big|_{n=n_t} = -\sum_{\tau=1}^{t+1} \varepsilon^{t+1-\tau} \sum_{k=1}^{K} \gamma_\tau(k)(I-A)^T \cdot (\Sigma_k^y)^{-1}(I-A)$$

In the analogous way, we can get the formulas of the channel parameter **h**. The practical algorithm execution steps in detail are described in [9].

4. EXPERIMENTAL RESULTS

4.1. Experiment Setting

To evaluate the effectiveness of the proposed algorithm, we perform a series of experiments on telephone quality and artificial contamination speaker-independent (SI) Mandarin speech recognition.

In order to obtain telephone quality speech materials for the acoustic model training, we utilize the Mandarin 863 speech database. The database was developed by Chinese National 863 Program for LVCSR. It contains about 70 hours' speech. The speech data are 16000Hz sampled and 16bit linear quantized. We disposed the database with three methods: 1) resample the database to 8000Hz and u-law quantization; 2) pass the database through the real PSTN network by Dialogic telephone cards plugged in PCs; 3) pass the database through GSM full rate (GSM FR 06.10) coder and decoder. All these three transcoded databases are used as training data for the acoustic model.

The acoustic features consist of energy, pitch, 12 mel-cepstral with delta and delta-delta features. The vocabulary of this task consists of more than 40K words. Tri-gram statistics are used for language modeling.

There are three test sets in our experiments named as TELTEST, GSMTEST, SIMUTEST, respectively. TELTEST is gathered through the PSTN network, GSMTEST is gathered through GSM-FR codec and SIMUTEST is derived from artificial exhibition contamination. Every subset contains about 240 continuous Mandarin sentences from 4 different speakers.

The compensation of noise speech is based on MMSE. Once the parameters of the distribution of the noise speech (y) are computed, an MMSE estimate is used to calculate the clean speech given the observed noisy speech:

$$\hat{x}_{MMSE} = E(x \mid y) = \int xp(x \mid y)dx = y - \sum_{k=0}^{K-1} P(k \mid y)(\mu_{y,k} - \mu_{x,k})$$

We also can use an alternative approach to correct means and variances of HMM [9] instead of performing the MMSE estimate of clean speech.

4.2 Comparison of the Actual Function and Our Approximation

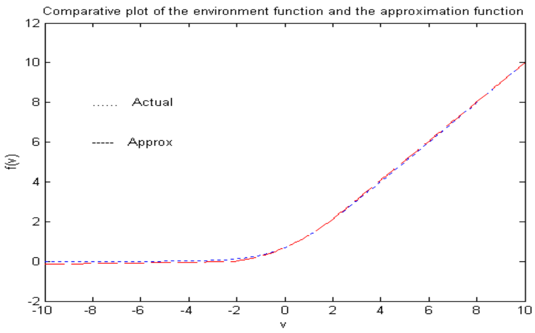

Fig.2 Comparative plot of the environment function and the approximation function

A comparative plot of the actual function and our approximation is shown in Fig.2. As can be seen from this figure, the approximation can't be distinguished from the actual function.

4.3 Comparison of VPP and Other Methods

We compare our VPP algorithm with other three widely used compensation algorithms.

1. Long term CMS (Cepstral Mean Subtraction over the duration of a whole Mandarin sentence).
2. SBR (Signal Bias Removal) [10].
3. VTS of order one [3].

The performance is measured by character error rate (CER) using NIST's SCLITE.

In Tab.1, we present the results from the experiment using speaker-independent Mandarin speech recognition in telecommunication environment. VPP1 denotes the compensation is applied to the batch EM, and, VPP2 to the recursive EM.

	Base-Line	CMS	SBR	VTS	VPP1	VPP2
TEL TEST	76.4	78.4	77.1	78.3	79.2	80.7
GSM TEST	75.9	76.3	78.7	79.8	80.4	81.1

Tab.1. Results on telephone quality speaker-independent Mandarin speech recognition

From Tab.1, VPP is the most effective one among these four compensation techniques. Compared with the baseline, VPP decreases the CER about 18% for TELTEST and 21.5% for GSMTEST. Especially for GSMTEST, where GSM channel is more nonlinear than fixed lines, VPP suits it best.

Fig.3 shows performance of VPP algorithm at several SNR's of SIMUTEST. We find that VPP obtains significantly lower WERs.

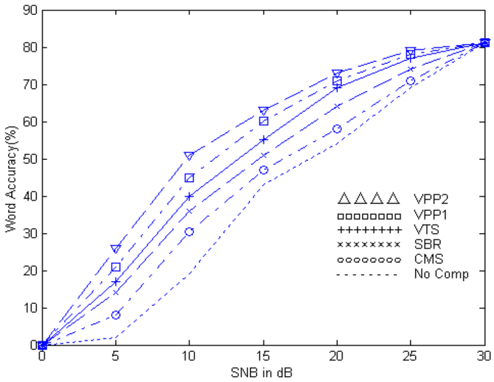

Fig.3 Performance of VPP algorithm at several SNR's

In the experiment of VPP1, the covariance matrices of the clean speech, the noisy speech and the additive noises are assumed to be diagonal in order to reduce the computational complexity. In the future, we will do some experiments on using the full matrices.

5. CONCLUSION

In this paper, we have presented a novel approximation–based approach to compensate for the effects of linear filter and background noise given the HMM parameters of clean speech. The algorithm is successfully applied in the batch EM and the recursive EM. The proposed experimental results showed that the approximation of environment function using the VPP algorithm is more precise than VTS. In the telephone network, the algorithm presented here provides significant improvement over the previous work in data compensation. Compared with other compensation approaches, VPP shows better adaptability to the variations introduced by channel and noise, then obtains about 18% character error rate decreasing in Mandarin telephone speech recognition.

6. REFERENCES

[1] J.T.Chien, H.C.Wang, and L.M.Lee, "Estimation of channel bias for telephone speech recognition," in Proc. ICSLP 1996, pp. 1840-1843.

[2] A. Acero, "Acoustical and Environmental Robustness in Automatic Speech recognition," Ph. D. Thesis, Department of Electrical and computer Engineering, Carnegie Mellon University, Sept. 1990.

[3] P.J.Moreno, "Speech Recognition in Noisy Environments," PhD thesis, Department of Electrical and computer Engineering, Carnegie Mellon University, April 1996.

[4] P.J.Moreno, B.Raj, E.B.Gouvea, and R.M.Stern, "Multivariate Gaussian Based Cepstral Normalization for Robust Speech Recognition". in Proc. ICASSP 1995, pp. 137-140.

[5] D. Y. Kim, C. K. Un, N. S. Kim, "Speech recognition in noisy environments using first-order vector Taylor series," Speech Communication, Vol. 24, No. 1, pp. 39-49, 1998

[6] N.S. Kim, "Nostationary environment compensation based on sequential estimation," IEEE Sig. Proc. Letters, Vol.5, pp.57-60, 1998

[7] V.Krishnamurthy and J.B.Morre, "Online estimation of hidden Markov model parameters based on the Kullback-Leibler information meature," IEEE Trans. Sig. Proc, Vol.41, pp. 2557-2573, 1993

[8] L.Deng, J.Droppo, and A.Acero, "Recursive noise estimation using iterative stochastic approximation for stereo-based robust speech recognition," ASRU 2001.

[9] A. Acero, L. Deng, T. Kristjansson, and J. Zhang, "Hmm Adaptation Using Vvector Taylor Series for Noisy Speech Recognition," in Proc. ICSLP 2000.

[10] Mazin G.Rahim, and Biing-Hwang Juang, "Signal Bias Removal by Maximum Likelihood Estimation for Robust Telephone Speech Recognition," IEEE Trans. On Speech and Audio Processing, Vol.4, No.1, pp.19-30, Jan, 1996.